Stacked Graphical Models for Efficient Inference in Markov Random Fields

Zhenzhen Kou
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA USA 15213

William W. Cohen
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA USA 15213

Abstract

In *collective classification*, classes are predicted simultaneously for a group of related instances, rather than predicting a class for each instance separately. Collective classification has been widely used for classification on relational datasets. However, the inference procedure used in collective classification usually requires many iterations and thus is expensive. We propose *stacked graphical learning*, a meta-learning scheme in which a base learner is augmented by expanding one instance's features with predictions on other related instances. Stacked graphical learning is efficient, especially during inference, capable of capturing dependencies easily, and can be constructed based on any kind of base learner. In experiments on 9 datasets, stacked graphical learning generally achieved comparable accuracy to other graphical models via much faster inference - stacked graphical learning can achieve similar performance in one or two iterations compared to Gibbs sampling which usually converge after 50 100 iterations .

1 Introduction

Traditional machine learning methods assume that instances are independent while in reality there are many relational datasets, such as hyperlinked webpages, scientific literatures with dependencies among citations, and social networks. The dependencies among data are complex. The instances could also have varying structures, for example, papers have different numbers of authors.

Collective classification has been widely used for classification on relational datasets. In collective classification, classes are predicted simultaneously for a group of related instances, rather than predicting a class for each instance separately. Recently there have been studies on relational graphical models, such as relational dependency networks [3], and relational Markov networks [4, 5]. Collective classification can be formulated as an inference problem over graphical models. Consider collective classification in the context of Markov random fields (MRFs). Inference in MRFs is intractible, in the general case. One common scheme for approximate inference is *Gibbs sampling*. Gibbs sampling for an MRF with parameters learned to maximize pseudo-likelihood is closely related to *conditional dependency networks* [**?**]. However, Gibbs sampling usually takes many iterations before converge and thus graphical models are usually expensive, especially when exact inference is infeasible.

We propose a meta-learning method, *stacked graphical learning*, for the learning and inference on relational data. In stacked graphical learning, a base learner is augmented by providing the predicted labels of related instances. That is, first, a base learner is applied to the training data to make predictions. Then we expand the features by adding the predictions of related examples into the feature vector. Finally the base learner is applied to the expanded feature set to obtain a stacked model.

One advantage of stacked graphical learning is that the inference is not iterative and thus very efficient. Experimental results show that compared to Gibbs sampling, stacked graphical learning can achieve similar performance in one or two iterations while Gibbs sampling usually converges after 50 100 iterations.

In stacked graphical learning, the dependencies among data can be captured easily using a *relational template* which finds the related instances given one example. Stacked graphical learning can be constructed based on any base learning algorithm, i.e., the base learner does not have to be a graphical model. Stacked graphical learning is easy to implement.

2 Algorithm

2.1 Stacked Graphical Learning We consider here collective classification tasks, in which the goal is to "collectively" classify some set of instances. In our notation, a collection of instances is a vector $\mathbf{x}$, and the labels for $\mathbf{x}$ are encoded as a parallel vector $\mathbf{y}$. Hence a collection is a pair $(\mathbf{x}, \mathbf{y})$ where each element of $\mathbf{x}$ is an instance–i.e., is itself a high-dimensional vector—and each element of $\mathbf{y}$ is a label from a small set $\mathcal{Y}$. In the paper we use upper case letters and their bold-faced equivalents, such as Y and $\mathbf{Y}$ for (vectors of) random variables.

We consider a model that captures the dependency by expanding the feature of an instance x_i with "predicted" labels for the related instances. We use the predicted label instead of true labels since during inference there is no way to get true labels. We use *relational template* C to pick up the related instances. A relational template is a procedure that finds all the instances related to a given example and returns their indices. For instance x_i, $C(x_i)$ retrieves the indices $i_1, ..., i_L$ of instances $x_{i_1}, ..., x_{i_L}$ that are related to x_i. Given predictions $\hat{\mathbf{y}}$ for a set of instances $\mathbf{x}$, $C(x_i, \hat{\mathbf{y}})$ returns the predictions on the related instances, i.e., $\hat{y}_{i_1}, ..., \hat{y}_{i_L}$. Since the relation between x_i and x_j might be one-to-many, for example, webpages link to different numbers of webpages, we allow aggregation functions to combine predictions on a set of related instances into a single feature.

One practical difficulty to obtain predictions for training examples is that, while learning methods produce reasonably well-calibrated probability estimates on unseen test data, their probability estimates on *training* data are biased. Thus, to obtain the "predictions" for training examples, we apply a cross-validation-like technique suggested by a meta-learning scheme, *stacking* [6]. The procedure to obtain the predictions for training examples is show in Figure 1.

Given a training set $D = \{(\mathbf{x}, \mathbf{y})\}$ and a base learner A, construct cross-validated predictions $\hat{\mathbf{y}}$ for $\mathbf{x} \in D$ as follows:

1. Split D into J equal-sized disjoint subsets $D_1 ... D_J$.
2. For $j = 1...J$, let $f_j = A(D - D_j)$.
3. For $\mathbf{x} \in D_j, \hat{\mathbf{y}} = f_j(\mathbf{x})$.

Figure 1: A cross-validation-like technique to obtain predictions for training examples

We wish to limit inference time by restricting the number of iterations performed by Gibbs to some small number K. To compensate for this severe restriction, we propose to allow the parameters used at each iteration of the sampling procedure to differ. This leads to the following variant of Gibbs sampling:

- for $i = 1 \ldots n$, pick $y_i^0 \sim \Pr(Y_i | \mathbf{X} = \mathbf{x}, \theta^0)$
- for $k = 1 \ldots K$
  - for $i = 1 \ldots n$,
    pick $y_i^k \sim \Pr(Y_i | \mathbf{Y}_{-i} = \mathbf{y}_{-i}^{k-1}, \mathbf{X} = \mathbf{x}, \theta^k)$

Notice that this is identical to standard Gibbs, except that it is based on $\theta^0, \theta^1, \ldots, \theta^K$. One approach would be learning $\theta^0, \theta^1, \ldots, \theta^K$ in a greedy fashion. We begin by setting θ^0 to maximize data likelihood, using a model restricted to the "non-relational" features. Then, for $k = 1, \ldots, K$, set θ^k to maximize likelihood of the data under the assumption that Q^{k-1} ($Q^t(\mathbf{Y}|\mathbf{X})$ is the distribution over $\mathbf{Y}^t$) is fixed, and k is the *final* iteration to be performed.

- Parameters: a relational template C and a cross-validation parameter J.
- Learning algorithm: Given a training set $D = \{(\mathbf{x}, \mathbf{y})\}$ and a base learner A:
  - Learn the local model, i.e., when $k = 0$:
    Return $f^0 = A(D^0)$. Please note that $D^0 = D, \mathbf{x}^0 = \mathbf{x}, \mathbf{y}^0 = \mathbf{y}$.
  - Learn the stacked models, for $k = 1...K$:
    1. Construct cross-validated predictions $\hat{\mathbf{y}}^{k-1}$ for $\mathbf{x} \in D$ as follows:
       (a) Split D into J equal-sized disjoint subsets $D_1 ... D_J$.
       (b) For $j = 1...J$, let $f_j^{k-1} = A(D^{k-1} - D_j^{k-1})$.
       (c) For $\mathbf{x} \in D_j, \hat{\mathbf{y}}^{k-1} = f_j^{k-1}(\mathbf{x}^{k-1})$.
    2. Construct an extended dataset $D^k = (\mathbf{x}^k, \mathbf{y})$ by converting each instance x_i to x_i^k as follows: $x_i^k = (x_i, C(x_i, \hat{\mathbf{y}}^{k-1}))$, where $C(x_i, \hat{\mathbf{y}}^{k-1})$ will return the predictions for examples related to x_i such that $x_i^k = (x_i, \hat{y}_{i_1}^{k-1}, ..., \hat{y}_{i_L}^{k-1})$.
    3. Return $f^k = A(D^k)$.
- Inference algorithm: given $\mathbf{x}$:
  1. $\hat{\mathbf{y}}^0 = f^0(\mathbf{x})$.

  For $k = 1...K$,
  2. Carry out Step 2 above to produce $\mathbf{x}^k$.
  3. $\mathbf{y}^k = f^k(\mathbf{x}^k)$.

  Return $\mathbf{y}^K$.

Figure 2: Stacked Graphical Learning and Inference

Finally we end up with the inference and learning methods of Figure 2 for collective classification. The relational template can be extended to include aggregation functions based on $\hat{\mathbf{y}}$ and x_i. We will demonstrate the use of aggregations in Section 3.

2.2 Stacked Graphical Learning: Sufficiency Study If the data is produced by Gibbs sampling

that converges geometrically, i.e., if $|\Pr^n(\mathbf{Y}|\mathbf{X}) - \Pr^{\inf}(\mathbf{Y}|\mathbf{X})| < (1-\epsilon)^n$ for all $\mathbf{X}$, $\mathbf{Y}$, n and initial distributions $\Pr^0$ (here $\Pr^k$ is the distribution after k iterations of Gibbs sampling), the stacking method will converge at least as fast, i.e., $|Q^n(\mathbf{Y}|\mathbf{X}) - \Pr^{\inf}(\mathbf{Y}|\mathbf{X})| < (1-\epsilon)^n$. This can be proven as follows: It is obvious for n=0 since the starting point is θ^0, same as Gibbs sampling that converges. If it is true for Q^{n-1}, by the convergence of the target chain, there is some θ^n that reduces the error by a factor of $1-\epsilon$ starting with Q^{n-1}, namely, θ^+, the parameters for Gibbs sampling. The stacking algorithm will take the θ^n that reduces error the most on Q^{n-1} on the data; in the limit, i.e., ignoring overfitting issues, this will reduce the error by at least $(1-\epsilon)^n$.

2.3 An alternative approach *[Comment : Introduce the further approximation: k-step-dependency stacking]* This approximation leads to an inhomogeneous chain that is only two steps long.

A further approximation is to set $\mathbf{y}^1$, $\mathbf{y}^2$, ..., $\mathbf{y}^K$ to all be equivalent, and all equal to $\mathbf{y}^0$. This approximation is conceptually useful, as it allows us to "unroll" the functions defined by θ^1, θ^2, ..., θ^K, and enumerate the features that contribute to each.

Let MB_i denote variables in the Markov blanket (set of related instances) for y_i, i.e., the set such that $\Pr(Y_i|\mathbf{Y}_{-i}, \mathbf{X}) = \Pr(Y_i|\mathrm{MB}_i, \mathbf{X})$. Let MB_i^2 be the set of all variables Y_ℓ such that $Y_\ell \in \mathrm{MB}_j$ for some $Y_j \in \mathrm{MB}_i$—i.e., MB_i^2 is the Markov blanket of the Markov blanket of Y_i. Likewise, let MB_i^k be the "order k Markov blanket of i", or set of all variables Y_ℓ such that $Y_\ell \in \mathrm{MB}_j$ for some $Y_j \in \mathrm{MB}_i^{k-1}$. Intuitively, MB_i^K is a diameter-K subgraph centered around Y_i—a sort of generalized "sliding window".

Letting $\tilde{Q}^k$ be approximation to Q^k that is obtained in this way, it is easy to see that

$$\begin{array}{rcl} \tilde{Q}^1(Y_i = y_i|\mathbf{x}) & \text{depends only on} & \mathbf{x}, y_i, \mathrm{MB}_i(\mathbf{y}^0) \\ \tilde{Q}^2(Y_i = y_i|\mathbf{x}) & \text{depends only on} & \mathbf{x}, y_i, \mathrm{MB}_i^2(\mathbf{y}^0) \\ \cdots & \cdots & \cdots \\ \tilde{Q}^K(Y_i = y_i|\mathbf{x}) & \text{depends only} & \mathbf{x}, y_i, \mathrm{MB}_i^K(\mathbf{y}^0) \end{array}$$

This suggests a new approximate learning method, in which we set θ^0 as in the greedy method above; construct, in some way, an expanded set of features that relate the value of y_i to the values of $\mathbf{x}$ and $\mathrm{MB}_i^K(\mathbf{y}^0)$; and finally find parameters $\tilde{\theta}$ that are MAP estimates for the dataset D using these features. This approximation leads to an inhomogeneous chain that is only two steps long.

In prior work [7], this sort of approximation was found to be effective for certain sequential classification problems—a special case of the collective classification task considered here in which the Markov network is a linear chain. However, there are reasons to believe that this "window" approximation will be less appropriate for general graphs. Consider a simple case, where the expanded feature set includes only a single edge between y_i and each $y_j \in \mathrm{MB}_i$, the Markov network has a maximum clique size of 2, and every $|\mathrm{MB}_i|$ is bounded by some small constant b. Let n be the number of parameters in the model θ^+. It is easy to see that the number of features used in this "unrolled" 2-step chain can grow rapidly with K: because $|\mathrm{MB}_i^K|$ can grow as b^K, the 2-step chain can have up to $n \cdot b^K$ parameters. This means that learning and evaluating the classifiers used in the second step of the chain will be expensive, and that the learner will be prone to overfit. While with linear chains, such as sequential classification problems, MB_i^K contains only $O(K)$ variables, which is probably why overfitting is less of a problem in [7].

Under the same assumptions, the method of Figure 2 will use only $n \cdot K$ parameters: even though $Q^K(Y_i = y_i|\mathbf{x})$ also depends on the values of $\mathbf{x}, y_i, \mathrm{MB}_i^K(\mathbf{y}^0)$, this dependency is "funnelled through" small Markov blankets involving variables $\mathbf{y}^1$, ..., $\mathbf{y}^{K-1}$. In the experiments, we will show that the method of Figure 2 is less liable to overfit than the two-step approximation.

3 Experimental Results

3.1 Datasets We evaluated stacked graphical learning on several classification problems. The first problem we studied is the task of text region detection in the system called the Subcellular Location Image Finder (SLIF) [8, 9]. SLIF is a system which extracts information from both figures and the associated captions in biological journal articles. Usually there are multiple *panels* (independently meaningful sub-figures) within one figure. Finding the text regions, i.e., the regions in panels containing their labels, is one important task in SLIF. The problem studied in this paper is to classify if the candidate regions found via image processing are text regions or not. The text region detection dataset contains candidate regions found in 1070 panels from 207 figures.

There are dependencies among the locations of candidate regions. Intuitively, if after image processing a candidate text region was found at the upper-left corner of panel B and two candidate regions were found in panel A, one located at the upper-left corner, another in the middle, it is more likely the candidate region at the upper-left of panel A is the real text region. We define the *neighbor* of a candidate text region to be the region located in the "same" position in adjacent

panels in the same figure and consider the neighbors on four directions, left, right, upper, and lower. We also consider the dependency among candidate regions within the same panel, called *competitors*. Figure 3 is an example figure in SLIF which demonstrates candidate regions, neighbors and competitors.

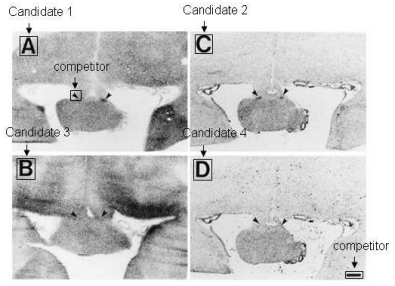

Figure 3: An example figure in SLIF

The relational template returns the predictions on one candidate region's neighbors and competitors. Since one candidate region can have several competitors from the same panel, we apply an EXISTS aggregator to the competitors, i.e., as long as there is one competitor which is predicted to be a text region, we assign 1 to the corresponding feature added during stacking.

In the real-world SLIF dataset, there are usually a few (usually 2∼10) panels in one figure. Thus the connected sub-graphs are relatively small. We generated a synthetic SLIF dataset with the same distributions over features and labels as the real-world data set, allowing more panels in one figure. We evaluated our approach on this synthetic dataset too. The synthetic dataset contains approximately 7500 candidate regions from 3200 panels in 200 figures.

The second problem is the document classification problem. We consider the webpage classification on WebKB dataset[10], which contains webpages from four computer science departments, the paper classification on Cora dataset and CiteSeer dataset. The webpages were manually labelled with one of the six categories: course, faculty, student, staff, research projects, or other. The data contains aproximately 3800 webpages and 8000 hyperlinks. The relational template applies the COUNT aggregator and returns the number of outgoing and incoming links in each category, given one webpage.

[Comment : describe the model the Cora dataset..... the CiteSeer dataset.....]

We use a maximum entropy learner as the base learner in stacked graphical learning for the first and second problem.

The third problem we study is the protein named entity extraction from Medline abstracts. We used three datasets to evaluate our method. The University of Texas, Austin dataset contains 748 labeled abstracts[1]; the GENIA dataset contains 2000 labeled abstracts[2]; and the YAPEX dataset contains 200 labeled abstracts[3].

We use conditional random fields (CRFs) as the base learner in stacked graphical learning for protein name extraction.

[Comment : We also evaluate on a person name extraction problem. the cspace data, the base learner]

[Comment : Vitor's data, base learner]

3.2 Performance of Stacked Graphical Learning To evaluate the effectiveness of stacked graphical learning, we compare four models. The first model is a competitive graphical model. For the first and second problem, we compare to relational dependency network (RDN) models [3]. The RDN model uses the same features as the stacked model, but learning via a pseudo-likelihood method and inference with Gibbs sampling. For name extraction (the third and fourth problem), we compare to stacked sequential model. *[Comment : Vitor's data .....]* The second model is a local model, i.e., the model trained with the base learner. The third model is the stacked graphical model. The fourth model is a probabilistic upper-bound (noted as ceiling model in Table 1) for the stacked graphical model, i.e., we use the stacked graphical model but allow true labels of related instances to be added during the feature extension. Table 1 shows the accuracy for each of the four models on two real-world datasets and the synthetic dataset. We used 5 fold cross validation unless for WebKB data we used 4 fold cross validation by departments. We use paired t-tests to access the significance of the accuracy. The t-tests compare the stacked graphical models with k=1 to each of the other three models. The null hypothesis is that there is no difference in the accuracy of the two models. The differences that are statistically significant at a $p < .05$ level are reported in the table with - or +.

On four of the tasks, the stacked graphical models improve the performance of the base learner significantly. Classification on WebKB data with the strong feature set is the only variant where stacking does not give a significant improvement. Yet in this case, there is no significant difference between the performance of the four models. On all the tasks, stacked graphical learning

[1] ftp://ftp.cs.utexas.edu/pub/mooney/bio-data/proteins.tar.gz

[2] http://www-tsujii.is.s.u-tokyo.ac.jp/genia/topics/Corpus/posintro.html

[3] http://www.sics.se/humle/projects/prothalt

Table 1: Evaluation on four models

	SLIF data	document classification			Vitor's	Cspace	Protein		
		WebKB	Cora	CiteSeer			UT	Yapex	Genia
RDNs	86.7	74.2							
Local model	77.2^{-}	58.3^{-}							
Stacked model (k=1)	90.1	73.2							
Stacked model (k=2)	90.1	72.1							
Ceiling for stacked model	96.3^{+}	73.6							

achieves statistically indistinguishable results to RDNs and on the WebKB tasks, stacked graphical learning achieves comparable results to the ceiling models. With one more iteration of stacking, usually there is no significant improvement of accuracy. On the WebKB task with the base feature set, the RDN model obtains an average accuracy of 74.2% while the ceiling model obtains an average accuracy of 73.6%. However, the difference is not statistically significant.

We notice that on the two classification problems, the base learner obtains a higher accuracy with the additional features. However, the differences in the accuracies of RDNs, stacked graphical models, and ceiling models on different feature sets are not statistically significant.

[Comment : results on protein name extraction, compared to standard CRFs and sequence stackedCRFs]

3.3 Convergence of Stacked Graphical Learning and Gibbs Sampling Figure 4 shows the convergence rate of stacking compared to Gibbs sampling on RDNs. The plots were generated using SLIF data and WebKB data with the basic feature set. We created the plots with a natural logarithm of k, where $ln(k) = -1$ corresponds to the initials where $k = 0$, and $ln(k) = 0$ corresponds to $k = 1$. We observe that stacked models (green curves) converge more quickly than Gibbs sampling and achieve a satisfactory performance much faster, even if the Gibbs sampling starts with same $\mathbf{y}^0$ as the corresponding stacked graphical models (red curves). Stacked graphical model can achieve significant improvement over the base learner after the first iteration. More iterations of stacking do not seem to be more helpful, with the performance keeping at the same level. We observe that Gibbs sampling converges to a same level after much more iterations and the convergence rate when k is small depends much on the starting points. We plot error bars along the curve for Gibbs sampling with random starting points. The error bars are calculated over 5 randomly initial samples, i.e., in each fold, Gibbs sampling is run 5 times with random initials. The standard deviation is calculated in each fold, and averaged to get the error bar.

The further approximation described in Section 3.3 suggests that a multi-stage stacking is related to a single-stage stacking with features from an order-k "window". Figure 4 shows that this is true when k is small, while the two-step approximation tends to overfit when the window size k grows (black curves).

[Comment : Fix the plot to make it look clearer. In addition to the plot showing the number of iterations, show the CPU time. Say something about the training, see if you can finish the experiments with online learner/piecewise CRFs, include them as future work, or part of this paper]

4 Conclusions

In this paper we presented stacked graphical learning, a meta-learning scheme in which a base learner is augmented by expanding one instance's features with predictions on other related instances. We showed that stacked graphical learning is a kind of short inhomogeneous Markov chains. Compared to other graphical models, stacked graphical learning is efficient, especially during inference. This property allows it to be very competitive in applications where an efficient inference algorithm becomes extremely important. The results on two classification problems indicate that classification with stacked graphical models can improve the performance of a base learner significantly and achieve accuracy competitive to other graphical models via much faster inference.

Cohen and Carvalho introduced stacked sequential learning in their paper [7]. In this paper, we extend the stacked sequential model to a more general case, where relational data is considered as the application, and demonstrate that stacked graphical models are short inhomogeneous Markov chains. McCallum and Sutton introduced parameter independence diagrams for introducing additional independence assumptions into pa-

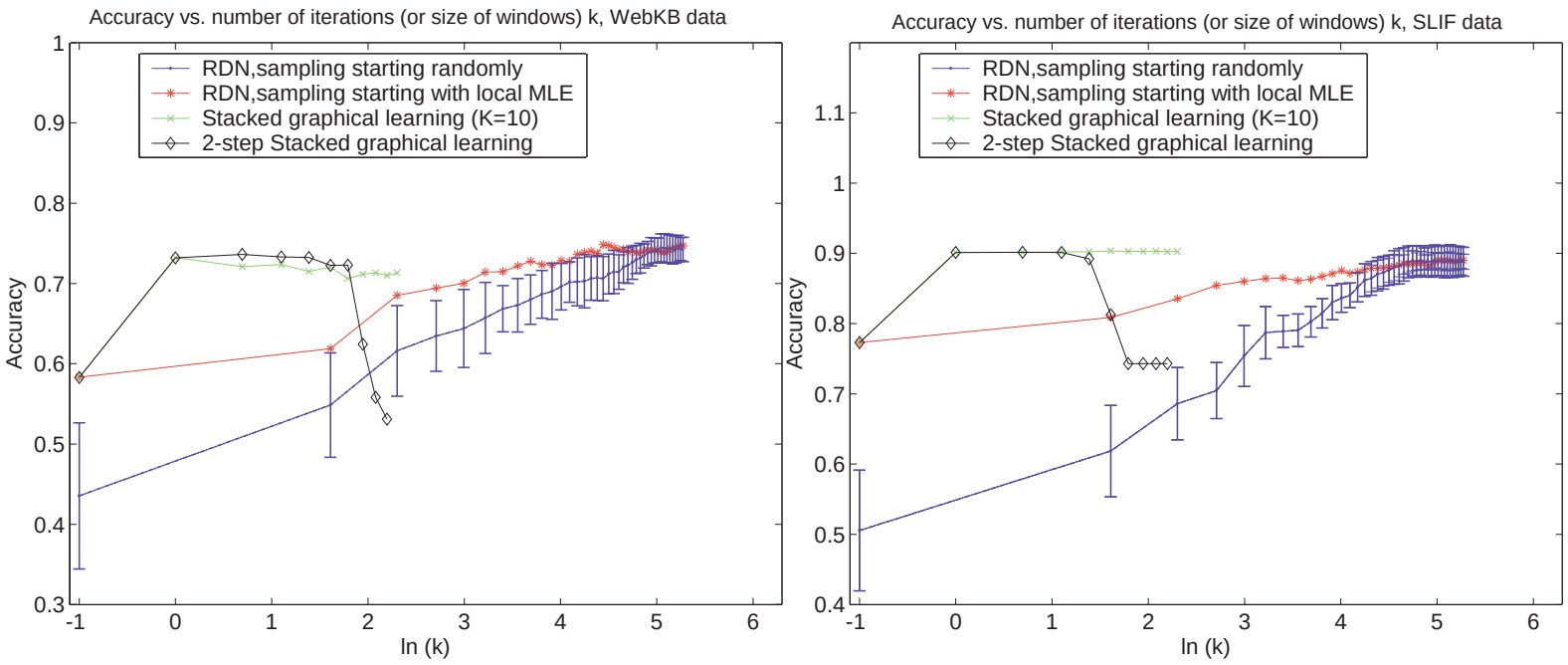

Figure 4: Convergence rate of stacking and Gibbs sampling

rameter estimation for efficient training of undirected graphical models[11]. Their method obtained a gain in accuracy via training in less than one-fifth the time. Our work is focusing on an approach which is efficient in inference.

Future work will compare stacked models to more graphical models such as relational Markov networks, and explore more on the relational template design and base learner selection. For example, integrating an online learning algorithm will enable fast training of stacked graphical models. Also there are more ways to design the relational template and expand the features during stacking. For example, when there are many features in the original feature set, simply "adding" the predictions might not work well. We are also considering the application to inter-related classification problems in an information extraction system.

References

[1] K. P. Murphy, *Bayes Net Toolbox for Matlab*, Computing Science and Statistics, 33(2201).

[2] J. Lafferty, A. McCallum and F. Pereira, *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*, Proceedings of the International Conference on Machine Learning (ICML-2001). Williams, MA 2001.

[3] D. Jensen and J. Neville , *Dependency Networks for Relational Data*, Proceedings of 4th IEEE International Conference on Data Mining (ICDM-04), Brighton, UK 2004.

[4] B. Taskar and P. Abbeel and D. Koller , *Discriminative Probabilistic Models for Relational Data*, Proceedings of Eighteenth Conference on Uncertainty in Artificial Intelligence (UAI02), Edmonton, Canada, 2002.

[5] R. Bunescu and R. J. Mooney, *Relational Markov Networks for Collective Information Extraction*, Proceedings of the ICML-2004 Workshop on Statistical Relational Learning (SRL-2004), Banff, Canada, 2004.

[6] D. H. Wolpert, *Stacked generalization*, Neural Networks, vol. 5, pp241–259, 1992.

[7] V. R. Carvalho and W. W. Cohen, *Stacked Sequential Learning*, Proceedings of Nineteenth International Joint Conferences on Artificial Intelligence, Edinburgh, Scotland, 2005.

[8] Z. Kou, W. W. Cohen and R. F. Murphy, *Extracting Information from Text and Images for Location Proteomics*, Proceedings of the BIOKDD 2003, Washington D.C., 2003.

[9] R. F. Murphy, Z. Kou, J. Hua, M. Joffe and W. W. Cohen, *Extracting and Structuring Subcellular Location Information from on-line Journal Articles: the Subcellular Location Image Finder*, Proceedings of the IASTED International Conference on Knowledge Sharing and Collaborative Engineering, St. Thomas, US Virgin Islands, 2004.

[10] M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam and S. Slattery, *Learning to Extract Symbolic Knowledge from*

the World Wide Web, Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98), Madison, WI, 1998.

[11] A. McCallum and C. Sutton, *Piecewise Training of Undirected Models*, Proceedings of 21st Conference on Uncertainty in Artificial Intelligence, 2005.